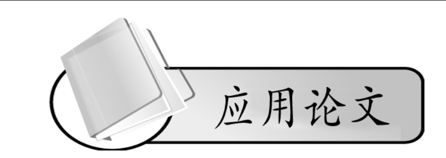


结合多种特征的 CRF 模型用于化学物质–疾病命名实体识别

隋明爽　崔　雷
(中国医科大学医学信息学院　沈阳 110122)

摘要: 【**目的**】建立结合多种特征的条件随机场模型, 探索从大型生物医学文本中同时自动提取化学物质和疾病实体的方法。【**方法**】结合命名实体识别特征, 包括词法特征、领域知识特征、词典匹配特征和无监督学习特征等, 比较不同特征对命名实体识别的效果, 并优化模型。【**结果**】CRF 模型纳入词法特征、词典匹配特征、无监督学习特征和部分领域知识特征, 化学物质识别准确率 97.33%、召回率 80.76%、F 值 88.27%, 疾病实体识别准确率为 84.20%、召回率为 81.96%、F 值为 83.07%。【**局限**】同时识别化学物质和疾病实体可能存在互相干扰, 删除的部分领域特征可能含有有用信息。【**结论**】本研究可为生物医学命名实体识别的特征选择提供参考, 同时仍需优化特征以获得更好的识别效果。

关键词: 命名实体识别　条件随机场　文本挖掘　无监督学习
分类号: TP391　G353

1　引　言

复杂的化学物质对疾病的作用机制错综复杂, 使人们对其相关药物的安全性更加警惕, 据统计, 化学物质与疾病的关系已成为 PubMed 数据库中用户检索最广泛的主题之一[1]。而由于临床试验的长期性、复杂性, 批准上市药物的副作用反馈机制的延迟, 导致难以早期预测到由化学物质导致的疾病相关信息。

与此同此, 生物医学文献爆炸性增长, 其中蕴含的化学物质–疾病直接与间接关系与漫长的临床试验相比无疑更为敏感。随着计算机和文本挖掘技术, 如自然语言处理等的发展, 使得从大型、非结构化的自由文本中识别和提取化学物质–疾病关联成为可能。本文旨在比较并结合多种特征, 建立条件随机场(Conditional Random Fields, CRF)模型, 以探索从大型生物医学文本中同时自动识别化学物质和疾病实体的方法。

2　研究现状

识别化学物质和疾病实体涉及到的关键技术主要是命名实体识别(Named Entity Recognization, NER), 即从生物医学数据中识别出化学物质和疾病实体, 主要思路是将识别任务转化为对文本基本单位的类别标注。其复杂之处在于, 现阶段药物和疾病命名实体数量的爆炸性增长、命名实体构词形式的多样性和低规律性, 以及命名规则(尤其是药物)不统一等[2]。

对于目前 NER 而言, 常用方法可分为基于规则、基于词典和基于机器学习等。

(1) 基于规则(模板): 通过规则大致描述使用的语法、句法、词汇、形态以及书写的特点长期形成的模式, 如对于化学物质实体来说, 其表达通常由大小写字母、数字、连字符(.和/)、希腊字母、罗马数字、

通讯作者: 崔雷, ORCID: 0000-0001-9479-8225, E-mail: lcui@mail.cmu.edu.cn。

引号、括号等字符组成。基于规则的 NER 系统通常依赖于由领域专家设计的规则，并通过正则表达式实现。如徐博等采用基于上下文模板的方法，从 PubMed 中构建了丰富的药物词典，不仅可识别出 DrugBank 中已有药名，甚至还能识别该库中没有的药物[3]。Tikk 等融合了基于规则的方法和条件随机场方法进行药物实体识别[4]。但因为对专家知识的依赖，这种类型的命名实体识别系统缺乏可扩展性和适应性。

(2) 基于词典：依赖于现有的词典识别自由文本中的命名实体，通常基于字符串匹配或字符串相似的算法，其性能取决于底层术语资源是否全面及算法性能。何林娜等运用基于词典和 CRF 相结合的方法，利用 PubMed 信息构建药物词典，并利用特征耦合泛化等方法对词典进行去噪，获得了较好的 F 值[5]。

(3) 基于机器学习：目前较流行的 NER 方法，对词典和规则的依赖性较小，可适用于不同领域，缺陷在于需要手动注释语料库。机器学习模型的性能取决于对于文本特征的辨别力以及算法[6]。其中，常用的方法有支持向量机模型、隐马尔可夫模型、最大熵模型和本文使用的 CRF 模型等。CRF 由 Lafferty 等[7]提出，结合了最大熵模型和隐马尔可夫模型的特点，是一种典型的判别式模型，已有研究证明其对于生物医学领域 NER 效果较好[8]。如 Lee 等使用改进的条件随机场算法提高疾病 NER 的水平[9]。

在实际使用中通常会有几种方法的结合，以获得更好的识别效果，如 Lowe 等提出使用结合语法和词典的方法进行化学物质 NER[10]，当前 tmChem[11]、DNorm[12]等工具和 NCBI 疾病语料库[13]的开发也为 NER 提供了便利。

3 研究思路与框架

本文参照 BioCreative V 大赛语料库[14]，结合当前 NER 领域常用的方法，构建一种结合词法特征、词典特征、领域知识特征和无监督学习特征的 CRF 模型，同时从生物医学文献中识别化学物质和疾病实体，通过反复调试对照，最终确定出识别效果最佳的 CRF 模型，如图 1 所示。流程实现主要依赖多种自然语言处理工具和 Perl 语言。

3.1 GENIA 预处理

GENIA Tagger 是专门针对生物医学文本的分析

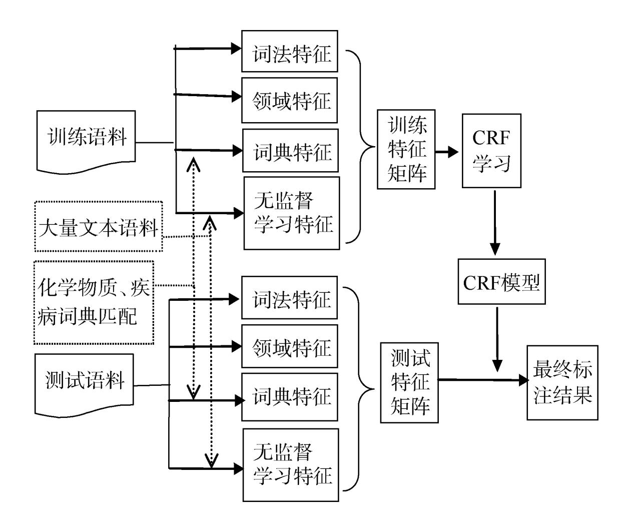


图 1　化学物质–疾病命名实体识别主要流程

器，可作为 NER 预处理的工具[15]。在 Linux 系统中运行 GENIA 程序，执行以下语句：./ geniaiagger <input> output。输入文件为切分好的句子(一句一行)，输出结果包括每个单词的词原形、词性、词块和对蛋白、DNA、RNA 和细胞系等实体的识别结果。

3.2 特征集构建

(1) 词法特征

①单词特征：以单词本身作为 NER 特征。

②词干、词性、词块特征：由 GENIA 运行结果获得。

③停用词特征：对每个词匹配停用词表，如果该词为停用词，则当前词停用词特征值为 Y，否则为 N。本文所用停用词表来自 tmChem[11]。

(2) 领域知识特征

①构词特征：对每个词进行正则表达匹配，以文本中当前词的构词形式是否符合大小写字母、数字、连字符(.和/)、希腊字母、罗马数字、引号、括号等特征为条件，如果符合则当前词的构词特征值为 Y，否则为 N，得到构词特征矩阵。

②高频词特征：高频词是指在化学物质和疾病命名实体中出现频率比较高的单词。本文通过如下步骤构建高频词列表[16]:

1)分别统计训练语料中标识为化学物质和疾病命名实体的单词，并记录单词在实体中的出现次数 CF；将单词在训练语料中出现的总次数记为 TF；

2)计算关键词特征信息的权重 Weight，其中 Weight=CF/TF×100%；

3)提取同时满足 TF≥100 和 Weight≥0.5；或同时满足 TF≥10 且 TF<100 和 Weight≥0.6；或同时满足 TF≥5 且 TF<10 和 Weight≥0.7；或同时满足 TF≥2 且 TF<5 和 Weight ≥0.8 的单词；最终分别得到化学物质和疾病高频词列表。以当前词是否在高频词列表中出现为特征，如果出现则当前词的关键词特征值为 Y，否则为 N。

除高频词特征外，以下词缀特征、词形特征、边界词特征、上下文特征列表构建也采用此方法。

③词缀特征：对每个单词分别取 3 个字符的前缀和后缀作为该单词的词缀特征。以前缀为例，分别统计训练语料中所有长度大于 5 的单词和实体的前 3 个字符组成训练语料前缀列表，统计各个前缀的出现次数 TF、CF。参照②获得化学物质前缀列表、后缀列表、疾病前缀列表、后缀列表。以当前词是否在词缀列表中出现为特征，如果出现则当前词的词缀特征值为 Y，否则为 N。

④词形特征：化学物质实体是一类特异性非常高的实体，其通常可能具有相同的词形。通过“AaX0”方式将大写字母替换为 A，小写字母替换为 a，数字替换为 0，其他字符替换为 X，构建词形特征。统计各种词形对应的单词数目 TF，和实体对应的各种词形的数目 CF，参照②获得化学物质词形列表。以当前词是否在列表中出现为特征，如果出现则当前词的词形特征值为 Y，否则为 N。

⑤边界词特征：边界词是指实体的第一个和最后一个单词。大部分实体是由多词组成，利用边界词信息可以提高边界识别能力，减少复合性实体的识别错误率。参照②构建化学物质左边界词和右边界词列表，疾病左边界词和右边界词列表。以当前词是否在列表中出现为特征，如果出现则当前词的边界词特征值为 Y，否则为 N。

⑥上下文特征：上下文信息是指实体前一个词和后一个词的单词信息，利用上下文信息可以提高基因实体边界识别能力。参照②构建化学物质上文和下文，疾病上文和下文列表。以当前词是否在列表中出现为特征，如果出现则当前词的上下文特征值为 Y，否则为 N。

⑦一元词和嵌套词特征：一元词指仅由一个单词构成的实体，嵌套词指不仅能独立作为一个命名实体，也能和其他单词组成复合命名实体，根据训练语料构建一元词特征和嵌套词词表；以当前词是否在列表中出现为特征，如果出现则当前词的词形特征值为 Y，否则为 N。

⑧tmChem①和 DNorm 特征：将 tmChem[11]和 DNorm[12]运行结果作为 CRF 特征之一：即以当前词是否被 tmChem[11]和 DNorm[12]标注为实体为特征，如果标注则当前词的特征值为 Y，否则为 N。tmChem 和 DNorm 是已有的、相对成熟的命名实体识别标注工具，本研究引入该特征旨在观察新 NER 模型与已有工具的比较，以及加入该特征能否改善 NER 结果。

(3) 词典匹配特征

构建化学物质和疾病词典，对语料词进行匹配，以当前词是否在词典单词列表中出现为特征，如果出现则当前词的词典单词特征值为 Y，否则为 N。

(4) 无监督学习特征

①词向量特征：采用 Word2Vec 开源工具②生成词向量[17-18]，将词表示为低维的、连续的、实值向量，是词表示的一种形式，输入样本量越大效果越好。本文利用 PubMed 数据库检索“C 大类 AND D 大类”的检索所有返回结果(检索式：“Chemicals and Drugs Category”[Mesh] AND “Diseases Category”[Mesh] AND hasabstract[text]，共 4 417 929 篇摘要)以及 BioCreaTive V 提供的训练集和测试集作为综合输入语料进行训练。本研究使用向量维度为 50，并采用 Wu 等[19]的方法将词向量矩阵简化为(+, –, 0)形式，公式如下：

$$\mathrm{MEAN}(j)^{+}=\frac{1}{N_j^{+}}\sum_{i=0}^{V}M_{ij} \qquad M_{ij}>0$$

$$\mathrm{MEAN}(j)^{-}=\frac{1}{N_j^{-}}\sum_{i=0}^{V}M_{ij} \qquad M_{ij}<0$$

$$M_{ij}^{*}=\begin{cases}+, \text{if} \quad M_{ij}>\mathrm{MEAN}(j)^{+}\\ -, \text{if} \quad M_{ij}>\mathrm{MEAN}(j)^{-}\\ 0, \text{otherwise}\end{cases}$$

其中，$\mathrm{MEAN}(j)^{+}$和 $\mathrm{MEAN}(j)^{-}$分别表示矩阵第 j 列的正均值和负均值。

②布朗聚类特征：Brown 等[20]提出一种基于词聚类的层次聚类算法③，按照从底层到顶层的顺序进行聚类。其输入是语料库中的词语，其输出是二进制元素构成的树结构。选取叶子节点的路径作为当前词的布朗聚类特征，用类似霍夫曼编码的方式对每个词进行编码，每个词都由一长串的二进制码构成。

③基于词向量的 K-means 聚类特征：在得到词向量的基础上，用 Word2vec 自带的 K-means 算法对词向量进行聚类，将相似度高的词聚在一起，本研究类别数取 256 类。

4 实证分析

4.1 数据源

BioCreative(Critical Assessment of Information Extraction in Biology，生物信息提取重要评估)是一项国际性大赛，致力于评价文本挖掘和信息提取系统在生物学领域的应用。2015 年间举办的 BioCreative V 任务包括疾病命名实体识别(Disease Named Entity Recognition and Normalization，DNER)和化学物质–疾病关系(Chemical-Disease Relations，CDR)提取两个子任务，可以作为本研究的语料库[13]，特别说明的是，该项目认为药物和化学物质之间是可相互转换的。

①https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/.

②https://code.google.com/p/word2vec/.

③https://github.com/percyliang/brown-cluster.

下载其训练集测试集[①]，从语料库中提取分别PMID、TI 和 AB 字段；句子切分后建立索引，转化为CRF 训练所需的标注模式。

4.2　词典构建

《医学主题词表》(Medical Subject Headings, MeSH)，是美国国立医学图书馆编制的权威性主题词表，是一部规范化的可扩充的动态性叙词表，可以用来进行NER 和标准化。从 NLM 网站[②]下载 MeSH 词表。

采用如下方法构建词表:

(1) 提取 MeSH 词表中所有属于 C 大类(Chemicals and Drugs Category)和 D 大类(Diseases Category)的主题词(MH 字段，MESH HEADING)及其入口词(ENTRY 字段)，并建立索引，使每种化学物质与其MeSH ID 一一对应。

(2) 考虑到 MeSH 词表中化学物质和疾病名称很大一部分包含在其增补概念记录(Supplementary Concept Records, SCRs)，其中没有可以明确表明属于类别的字段，笔者制定如下规则从补充记录中提取化学物质和疾病：如果该记录的 HM 字段(Heading Mapped-to)指向 MeSH 词表中的 C 大类或 D 大类物质，则认为该记录表示一种化学物质或疾病，同样也编制相应索引，由此获得化学物质和疾病两个词表。

4.3　特征集列表构建

从训练语料中获得化学物质和疾病高频词列表、词缀列表、词形列表、左右边界词列表、上下文列表。以化学物质高频词列表为例，如表 1 所示:

表 1　化学物质高频词列表(部分)

单词	实体中出现频次(CF)	语料中出现频次(TF)	Weight=CF/TF
pirenzepine	4	4	4/4
dapsone	3	6	3/6
11-deoxycortisol	2	2	2/2
creatine	10	10	10/10
amphetamine	23	30	23/30
sulindac	6	6	6/6
ribavirin	15	15	15/15
AX	10	13	10/13

4.4　特征矩阵

参照实验方案设计，分别对特征集列表进行匹配，获得特征矩阵形式如表 2 所示，本文标注方法采用“IBO”模式，其中 I(Inside)表示当前词是命名实体中的一个词，O(Outside)表示当前词不是命名实体，B(Beginning)表示当前词是命名实体的首词。

表 2　特征矩阵(部分)

词	词法特征		领域特征			无监督学习特征				词典特征		标注结果
Butyrylcholinesterase	NN	B-NP	N	...	N	0	...	108	1110010000000000000000000000	Y	N	O
gene	NN	I-NP	N	...	N	-	...	53	11100100000000000000000000000000	N	N	O
mutations	NNS	I-NP	N	...	N	-	...	43	11100100000000000000000000000000000000000 00	N	N	O
in	IN	B-PP	N	...	N	0	...	117	0000000000000000000000000	N	N	O
patients	NNS	B-NP	N	...	N	-	...	132	1110010000000000000	N	N	O
with	IN	B-PP	N	...	N	0	...	158	1110010000000000000	N	N	O
prolonged	JJ	B-NP	N	...	N	0	...	25	111001000000000000	N	N	O
apnea	NN	I-NP	N	...	Y	+	...	178	11100100000000000000000000000000	N	Y	B-Diseases
after	IN	B-PP	N	...	N	0	...	25	11100100000000000000000000000000	N	N	O
succinylcholine	NN	B-NP	Y	...	N	0	...	164	111001000000000000000	Y	N	B-Chemical
for	IN	B-PP	N	...	N	-	...	134	1110010000000000000000000000000	N	N	O
electroconvulsive	JJ	B-NP	N	...	N	-	...	68	1110010000000000000	N	N	O
therapy	NN	I-NP	N	...	N	-	...	62	11100100000000000000000000000000	N	N	O
.	.	O	N	...	N	0	...	117	11100100000000000 11100101	N	N	O

①http://www.biocreative.org/.

②https://www.nlm.nih.gov/mesh/filelist.html.

4.5　执行 CRF 模型

本文采用 CRFs++ −0.58 开源工具包[①]构建化学物质–疾病命名实体识别模型。CRF++通过定义特征模板来提取训练语料中的特征，以此实现对训练语料进行学习。特征模板文件中的每一行代表 CRF++的特征提取模式，其中%x[row, col]表示输入数据中的一个 token, row 和 col 表示相对的行偏移与列偏移(见表 3)。使用的 CRF++模板均为 Unigram 类型，调整特征模板以获得最好的 NER 结果。

表 3　CRF++ 特征模板(部分)

Unigram
U1: %x[-2, 0]
U2: %x[-1, 0]
U3: %x[0, 0]
U4: %x[1, 0]
U5: %x[2, 0]
U6: %x[-2, 0]/%x[-1, 0]
U7: %x[-1, 0]/%x[0, 0]
U8: %x[0, 0]/%x[1, 0]
U9: %x[1, 0]/%x[2, 0]
…

5　实验结果

5.1　数据处理结果

本研究所使用训练集和测试集均包含 500 篇 PubMed 摘要，各包含 111 990 和 116 840 个单词，分别包含 5 203、5 385 个化学物质实体和 4 182、4 424 个疾病实体。MeSH 词表处理后得到包含 578 475 个化学物质术语的化学词典和包含 195 151 个疾病术语的疾病词典。

5.2　CRF 特征模板和 NER 结果

经调试比较，最终使用的模板共包含 21 个特征，其中 5 个化学特征，5 个疾病特征，11 个化学物质疾病公用特征；单词、词原形、词块等 8 个特征使用上下文窗口为 5 的模板。

不同 NER 模型识别结果如表 4 所示，可见只使用基于词典匹配结果准确率和 F 值较低，但召回率较高；加入词法特征后，化学物质和疾病实体的识别 F 值各自提高了 10%左右，准确率得到较大提高，而召回率都有所下降；加入领域知识特征后，化学物质 NER 虽然准确率有所升高，召回率继续较大幅度下降，导致 F 值随之下降, NER 效果变差，经过反复试验在最终的模型中去掉除了化学物质上下文和 tmChem 结果以外的领域知识特征，相比之下，疾病 NER 的准确率和召回率均有所上升，对比验证后，发现保留疾病实体上下文和 DNorm 结果特征结果最佳；加入无监督学习特征之后，两种 NER 效果均有所提高，最终化学物质 NER 的 F 值达到 88.27%，疾病 NER 的 F 值达到 83.07%，总体准确率、召回率和 F 值为 90.64%、81.32%、85.73%。

表 4　化学物质和疾病 NER 结果

特征		准确率(%)	召回率(%)	F 值(%)
词典匹配	Chemical	64.07	83.73	72.59
	Disease	59.09	82.41	68.83
词法+词典匹配	Chemical	91.07	74.77	82.12
	Disease	87.51	65.76	75.09
领域+词法+词典匹配	Chemical	96.94	59.36	73.63
	Disease	85.67	73.55	79.15
调整后领域+词法+词典匹配	Chemical	97.15	80.35	87.96
	Disease	85.10	79.61	82.26
调整后领域+词法+词典匹配+无监督学习	Chemical	97.33	80.76	88.27
	Disease	84.20	81.96	83.07

此外，在加入已有工具 tmChem 和 DNorm 之前，模型可达到的化学物质和疾病识别效果 F 值分别为 86.45%、80.13%，高于 DNorm(78.2%)和 tmChem (83.6%)，且在加入 tmChem 和 DNorm 特征后，模型总体识别效果 F 值约提高 3%。

但是可能由于同时识别化学物质和疾病实体的原因，导致单独疾病 NER 结果不甚理想，在 BioCreative V 所有参赛队伍中准确率、召回率、F 值分别排名 7, 4, 7(最高值分别为 90.53%、86.17%、86.46%；平均值 78.99%、74.81%、76.03%)。

对于识别错误的单词进行分析，可将主要错误类型分为:

(1) 错误标记的实体：即非实体被标记为实体，如 high Adherence group;

(2) 未识别出的实体：如复杂的化学实体 3,

①http://crfpp.sourceforge.net/.

4-methylenedioxymethamphetamine，不常见连词符实体 piperacillin/tazobactam 和缩写实体，如 iAs(inorganic Arsenic 缩写)；

(3) 边界识别错误的实体：即不完整或超出边界的实体，如 acute myocardial ischemia，会把程度词识别出来。

其原因可能是加入领域知识特征后，模型性能有所下降，故在最终模型中删除了构词特征、词形特征和边界词特征等可能对于识别上述类型实体有重要作用的特征。

6 结　语

本文在前人的研究基础上，对结合多种特征的 CRF 模型的识别结果进行探索，构建了一种同时从生物医学文献中识别化学物质和疾病实体的模型。所得模型涵盖了多种当前流行的特征，包括词法特征、领域知识特征、词典匹配特征和无监督学习特征，实验结果表明，领域知识特征中高频词特征、构词特征、词形特征和边界词特征对于化学物质 NER 表现不佳，最终模型中仅保留了领域知识的上下文和 tmChem 特征，但这一策略牺牲了部分实体的识别效果。

进一步研究方向将针对识别错误的实体类型进行后续处理，如将括号内的缩写词、连词符(and/or/-)连接词等定义为特殊特征，并对程度词进行限制。此外，对于无法准确识别的复杂化学物质实体，可通过完善词典和加强词形、构词特征权重加以识别。

本研究尚不完善，可为后续化学物质–疾病 NER 模型的特征确定提供参考，后期拟在此工作基础上开发关系提取算法，通过句法分析及机器学习算法提取化学物质–疾病实体，结合其语义关联整合为完整的化学物质–语义关系–疾病对，最终开发出从生物医学文本中自动识别和提取化学物质和疾病实体及其相互关系的平台。

参考文献：

[1] Wei C H, Peng Y, Leaman R. et al. Overview of the BioCreative V Chemical Disease Relation (CDR) Task[C]. In: Proceedings of the 5th BioCreative Challenge Evaluation Workshop. 2015.

[2] 隋明爽，崔雷．用文本挖掘方法发现药物的副作用[J]．中华医学图书情报杂志，2015，24 (11)：67-72．(Sui Mingshuang, Cui Lei. Detection of Drug Adverse Effects by Text-mining[J]. Chinese Journal of Medical Library and Information Science, 2015, 24(11): 67-72.)

[3] 徐博，林鸿飞，杨志豪．基于模板抽取和丰富特征的药名词典生成[C]．见：第五届全国信息检索学术会议论文集.2009. (Xu Bo, Lin Hongfei, Yang Zhihao. Generating a Drug Name Dictionary Based on Pattern Extraction and Rich Feature Sets[C]. In: Proceedings of the 5th China Conference on Information Retrieval. 2009.)

[4] Tikk D, Solt L. Improving Textual Medication Extraction Using Combined Conditional Random Fields and Rule-based Systems[J]. Journal of the American Medical Informatics Association, 2010, 17(5): 540-544.

[5] 何林娜，杨志豪，林鸿飞，等．基于特征耦合泛化的药名实体识别[J]．中文信息学报，2014, 28(2): 72-77. (He Linna, Yang Zhihao, Lin Hongfei, et al. Drug Name Entity Recognition Based on Feature Coupling Generalization [J]. Journal of Chinese Information Processing, 2014, 28(2): 72-77.)

[6] Krauthammer M, Nenadic G. Term Identification in the Biomedical Literature [J]. Journal of Biomedical Informatics, 2004, 37(6): 512-526.

[7] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. In: Proceedings of the 2002 International Conference on Machine Learning. 2002.

[8] Chowdhury Md F M, Lavelli A. Disease Mention Recognition with Specific Features [C]. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010.

[9] Lee H C, Hsu Y Y, Kao H Y. An Enhanced CRF-based System for Disease Name Entity Recognition and Normalization on BioCreative V DNER Task [C]. In: Proceedings of the 5th BioCreative Challenge Evaluation Workshop. 2015.

[10] Lowe D M, Sayle R A. LeadMine: A Grammar and Dictionary Driven Approach to Entity Recognition [J]. Journal of Cheminformatics, 2015, 7(S1): 1-9.

[11] Leaman R, Wei C H, Lu Z. tmChem: A High Performance Approach for Chemical Named Entity Recognition and Normalization [J]. Journal of Cheminformatics, 2015, 7(S1): 1-10.

[12] Leaman R, Islamaj Dogan R, Lu Z. DNorm: Disease Name Normalization with Pairwise Learning to Rank [J]. Bioinformatics, 2013, 29(22): 2909-2917.

[13] Doğan R I, Leaman R, Lu Z. NCBI Disease Corpus: A

Resource for Disease Name Recognition and Concept Normalization [J]. Journal of Biomedical Informatics, 2014, 47(2): 1-10.

[14] Li J, Sun Y, Johnson R J, et al. Annotating Chemicals, Diseases and Their Interactions in Biomedical Literature [C]. In: Proceedings of the 5th BioCreative Challenge Evaluation Workshop. 2015.

[15] Kim J D, Ohta T, Tateisi Y, et al. GENIA Corpus--Semantically Annotated Corpus for Bio-textmining[J]. Bioinformatics, 2003, 19(S1): 180-182.

[16] 夏光辉. 基于词典与机器学习的基因命名实体识别机制研究[D]. 北京: 北京协和医学院, 2013. (Xia Guanghui. The Research of Gene Name Entity Recognition Mechanism by Combining Dictionary Method and Machine Learning Method [D]. Beijing: Peking Union Medical College, 2013.)

[17] Zhang Y, Xu J, Chen H, et al. Chemical Named Entity Recognition in Patents by Domain Knowledge and Unsupervised Feature Learning [J/OL]. The Journal of Biological Databases and Curation [2016-06-10]. http://database.oxfordjournals.org/content/2016/baw049.

[18] 何红磊. 基于词表示方法的生物医学命名实体识别[D]. 大连: 大连理工大学, 2015. (He Honglei. Research of Word Representations on Biomedical Named Entity Recognition [D]. Dalian: Dalian University of Technology, 2015.)

[19] Wu Y, Xu J, Jiang M, et al. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text [C]. In: Proceedings of the 2015 AMIA Annual Symposium. 2015.

[20] Brown P F, Desouza P V, Mercer R L, et al. Class-based N-gram Models of Natural Language [J]. Computational Linguistics, 1992, 18(4): 467-479.

作者贡献声明:

隋明爽: 设计研究方案, 进行实验, 采集、清洗和分析数据, 论文起草;

崔雷: 提出研究思路, 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: suims1107@163.com。

[1] 隋明爽. Allchemical.txt; alldisease.txt. 化学物质和疾病词表.

[2] 隋明爽. corpus.pl. 数据处理及词典匹配算法.

[3] 隋明爽. feature.zip. 各特征构建算法.

[4] 隋明爽. CRF model.zip. 最终训练集、测试集和 CRF 模型.

收稿日期: 2016-06-24
收修改稿日期: 2016-07-19

Extracting Chemical and Disease Named Entities with Multiple-Feature CRF Model

Sui Mingshuang　Cui Lei

(School of Medical Informatics, China Medical University, Shenyang 110122, China)

Abstract: [**Objective**] This study aims to build a CRF model with multiple features, which could automatically extract chemical and disease named entities from biomedical documents. [**Methods**] We compared the performance of popular named entity recognition features, including lexical features, domain knowledge features, dictionary matching features as well as unsupervised learning features, and then optimized the new model. [**Results**] We built the final CRF model with lexical features, dictionary matching features, unsupervised learning features and part of the domain knowledge features. The precision, recall, and F-score for chemical entities identification tasks were 97.33%, 80.76%, and 88.27, respectively. For disease entities, they were 84.20%, 81.96%, and 83.07%, respectively. [**Limitations**] Chemical and disease entities may interfere with each other while being identified simultaneously. The deleted domain knowledge features may contain valuable information. [**Conclusions**] This study proposed a new method to identify biomedical named entities, which could be further improved.

Keywords: Named entity recognition　CRF　Text mining　Unsupervised learning